Analisi Esplorativa

ELABORAZIONE DEI DATI SCIENTIFICI

[262-009]

Autore:

Testoni Manuel (176931)

October 18, 2025

Contents

1

Introduzione

1.1 Contesto del dataset

Il dataset in questione contiene i dati relativi agli olii d'oliva provenienti da varie regioni di italia. Sono presenti all'interno 382 campioni, ognuno dei quali presenta 8 feature:

- Sigla di provenienza
- Acido Grasso Palmitico
- Acido Grasso Palmitoleico
- Acido Grasso Stearico
- Acido Grasso Oleico
- Acido Grasso Linoleico
- Acido Grasso Eicosanoico
- Acido Grasso Linolenico

  Un'osservazione preliminare che si può fare a partire dal dataset è che questo non è bilanciato, ovvero vediamo come il numero di campioni per ogni regione non sia il medesimo.

  Le regioni di provenienza dell'olio sono: **SA** (sud puglia); **NA** (nord puglia); **WL** (liguria ovest); **EL** (liguria est); **U** (Umbria)

1.2 Obiettivi dell'analisi

Il compito di questo studio sarà quello di andare a applicare le nozioni apprese a lezione al fine di analizzare le distribuzioni, i trend e le caratteristiche dei campioni di questo dataset con scopo quello di trarne nuova conoscenza.

1.3 Metodologia

Per applicare le varie metodologie discusse in aula sono state implementate soluzioni in **Matlab** [1]. Più nello specifico è stato realizzato uno script che, a partire dal caricamento del dataset, ha generato una serie di grafici che commenteremo nel corso dello studio.

2

Esplorazione dei dati

2.1 Importazione e comprensione del dataset

Oltre ad osservare che il dataset non è bilanciato come detto in precedenza, siamo anche in grado di dire che non vi sono valori mancanti o anomali.

2.2 Statistiche descrittive

É stata condotta un analisi preliminare per avere come riferimento i range e la media delle variabili. Nella tabella qua sotto è possibile osservare questi parametri raggruppati per ogni singolo acido grasso.

Table 2.1: Statistiche descrittive per ciascun acido grasso.

Acido grasso	Minimo	Massimo	Media
Palmitico	6.10	17.53	12.526
Palmitoleico	0.15	2.80	1.363
Stearico	1.52	3.50	2.204
Oleico	63.00	84.10	73.193
Linoleico	5.10	14.62	9.625
Eicosanoico	0.00	0.70	0.301
Linolenico	0.00	1.02	0.518

3

Visualizzazioni e Analisi

All'interno di questo capitolo andiamo ad analizzare figura per figura tutti i grafici che abbiamo prodotto all'interno della nostra Exploratory Analysis.

3.1 Line plot (Parallel Coordinates)

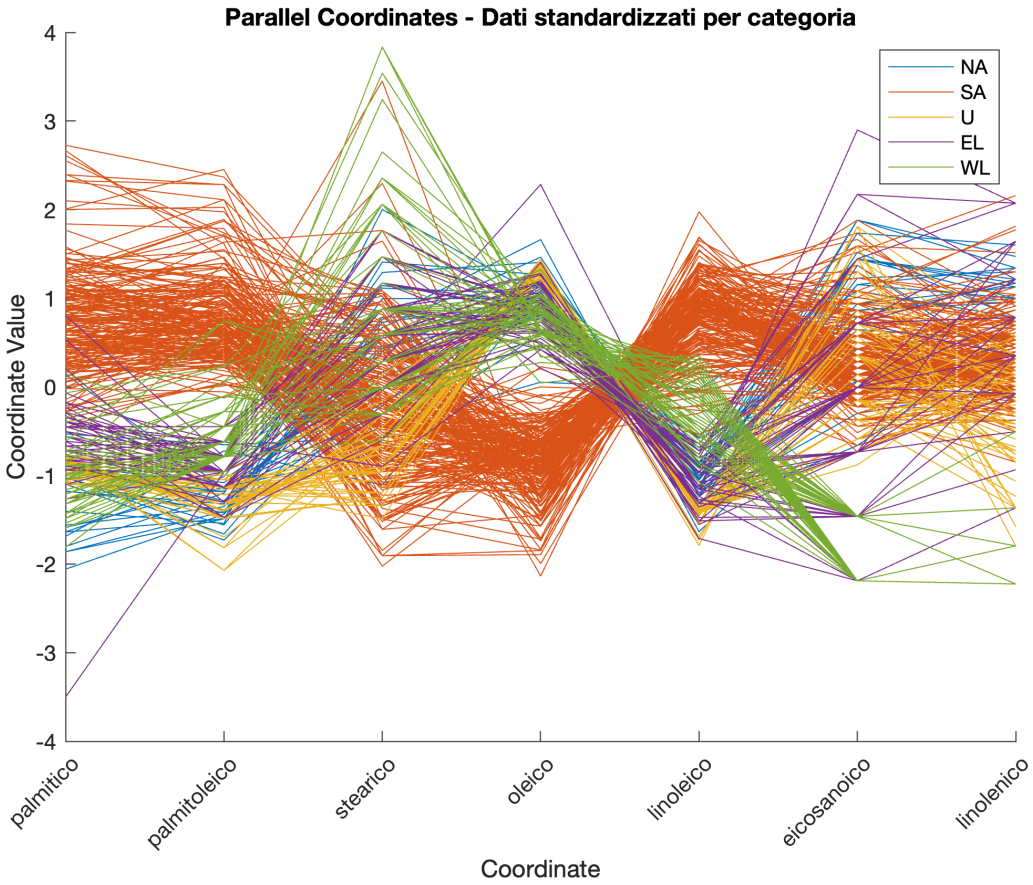


Figure 3.1: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.

Osservando il grafico in figura riusciamo, ancora una volta, a decretare che il dataset non è bilanciato in quanto i campioni "arancioni" quindi appartentni alla categoria **SA** sono di gran lunga maggiori rispetto agli altri.
Possiamo inoltre notare come prendendo le due variabili acidi grassi "oleico" e "linoleico" vi sia un comportamento inverso tra le varie regioni e **SA**. Infatti possiamo dire che vi è un calo dell'acido grasso "oleico" e un aumento dell'acido grasso "linoleico" per il Sud della Puglia, mentre per le altre regioni vale il contrario. In generale l'unico andamento che si discosta di più per tutti gli acidi grassi è proprio quello della Puglia, gli olii delle varie regioni sembrano essere abbastanza allineati.

3.2 Gplotmatrix

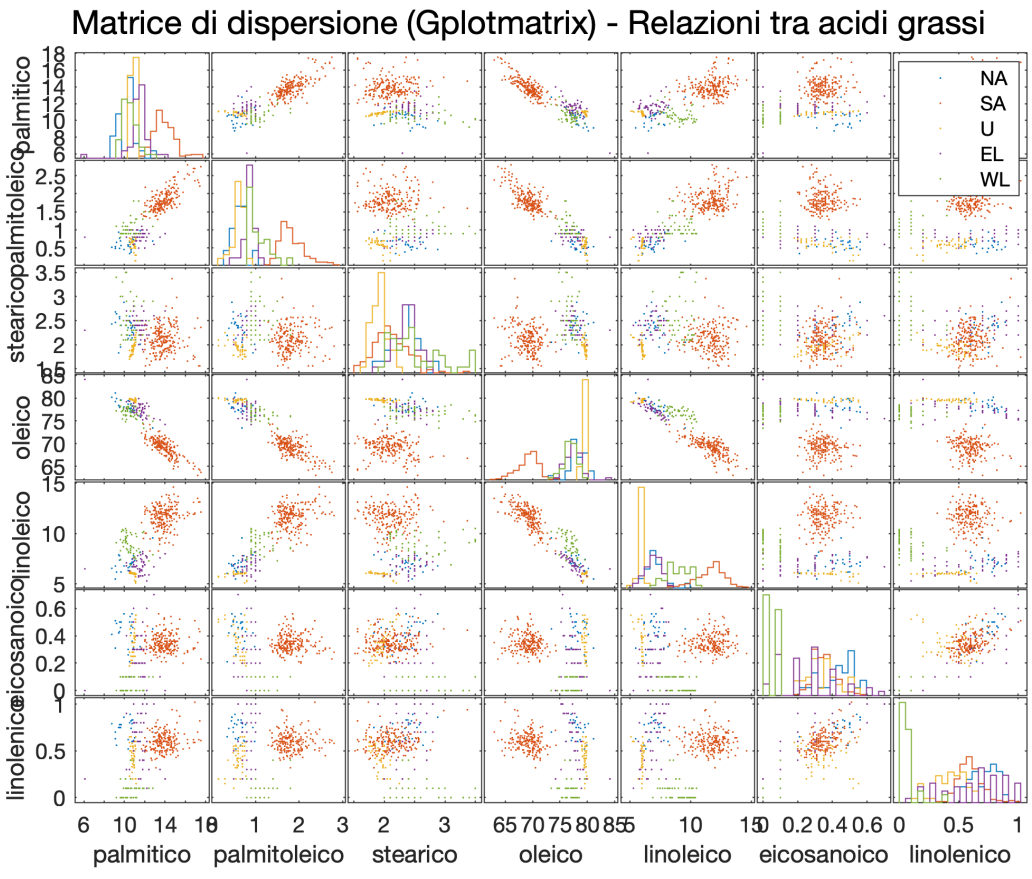


Figure 3.2: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.
La prima cosa che possiamo dire guardando questa matrice di dispersione è che i valori di NA per ogni acido sono sempre distaccati dagli altri. I pochi casi in cui vengono raggiunti da acidi di altre rgioni è perchè questi sono valori anomali. Tra gli acidi palmitoleico e oleico vi è un trend lineare inverso, le due variabili non sono quindi dipendenti. Vi è inoltre un trend inverso tra gli acidi palmitico e oleico. Un trend lineare è invece individuabile per gli acidi oleico e linoleico. Un ulteriore trend lineare è tra le variabili palmitico e palmitoleico. Due variabili che permettono di distinguere gli olii provenienti da WL sono l'acido eicosanoico e linolenico che in genere hanno valori bassi.

3.3 Analisi diagonale ed extradiagonale

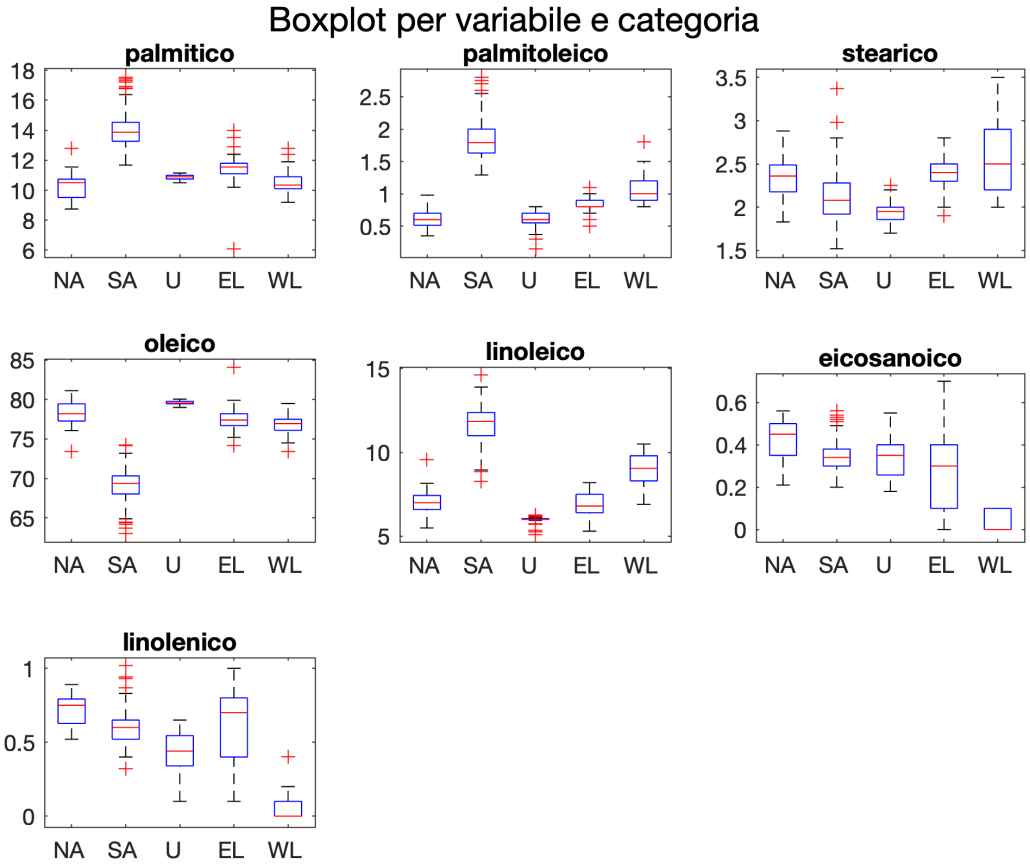


Figure 3.3: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.

Per quanto riguarda questo grafico facciamo commenti riguardo ad ogni singolo acido grasso:

- Palmitico: NA, EL e WL presentano strutture a gruppi per questo acido grasso e EL si distingue per un buon numero di campioni di campioni leggermente al di fuori della distribuzione dei dati e un outlier, quindi un campione totalmente al di fuori dalla distribuzione. Per SA e U abbiamo una linea della mediana nel mezzo quindi possiamo dire che non emergono strutture a gruppi.

- Palmitoleico: la distribuzione non è simmetrica per U, EL e WL, che presentano anche qualche campione al di fuori della distribuzione. Per quanto riguarda NA e SA invece abbiamo uan distribuzione simmetrica e SA presenta svariati campioni al di fuori della distribuzione.

- Stearico: in questo boxplot è ancora più evidente come queste variabili abbiano variabilità e valori diversi. Notiamo però che questo acido grasso per tutte le regioni la linea della mediana quasi a metà quindi non si evidenziano particoalri strutture a gruppi.

- Oleico: possiamo dire qui invece che U ha una variabilità minuscola senza presentare gruppi che fuoriescono dalla distribuzione. Per quanto riguarda invece NA, SA, EL, WL possiamo dire che la linea della mediana è quasi sempre sulla metà del box, quindi non si evincono strutture a gruppi.

- Linoleico: in generale la variabilità per questo acido gratto nelle varie regioni non è elevata, abbiamo che Na e WL non presentano strutture a gruppi, mentre EL e SA sì. Il box plot di U ha una variabilità così piccola che non si riesce nemmeno a rappresentare bene in scala. Presenta inoltre un sacco di valori al di fuori della distribuzione e sicuramente almeno 1 outlier.

- Eicosanoico: per questo acido grasso si nota facilmente che per WL la mediana corrisponde al valore del primo quartile. Mentre SA ha precisamente la linea della mediana verso metà il che suggerisce non avere strutture a gruppi. Ha inoltre qualche campione che non rispetta la distribuzione normale dei dati. NA, U, EL invece suggeriscono la presenza di strutture a gruppi avendo la linea della mediana tutti e tre spostata verso il terzo quartile.

- Linolenico: Quanto detto sopra vale anche per questo acido grasso se non fosse che rispetto a prima, anche U ha la mediana precisamente a metà.

3.4 Scatterhist per plot significativi

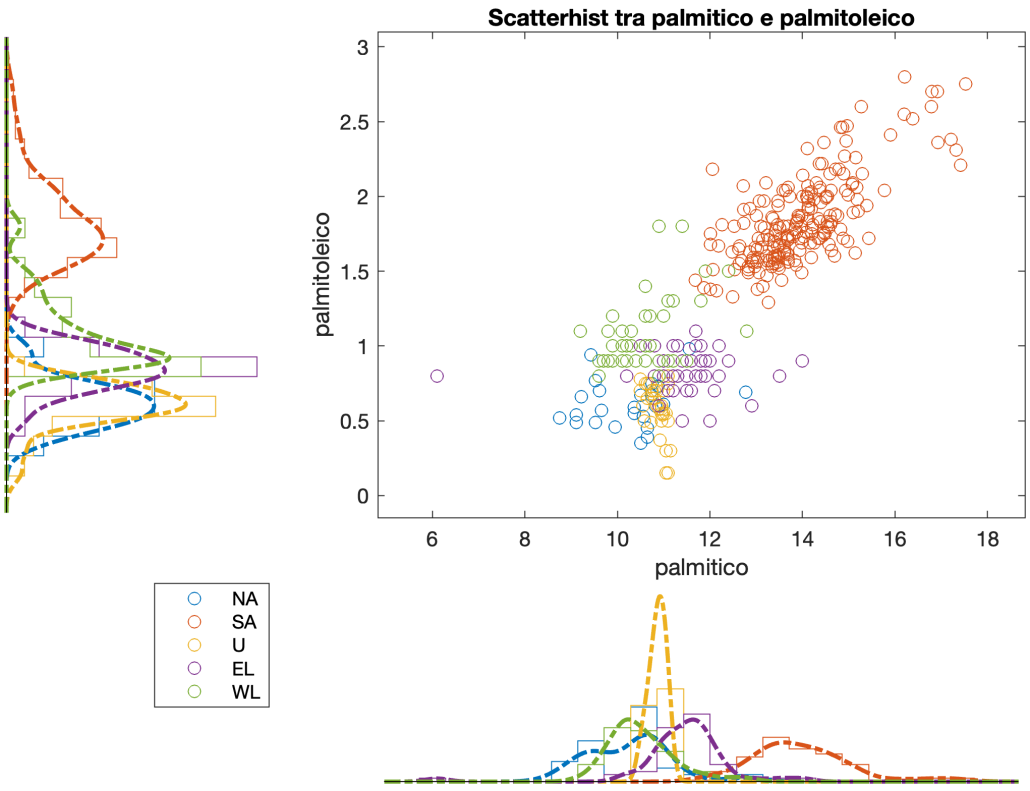


Figure 3.4: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.

In figura è rappresentato lo Scatterhist tra Acido Palmitoleico e Palmitico. Come possiamo vedere è presente un trend lineare, ovvero all'aumentare dell'acido palmitico si nota un aumento dell'aicdo grasso palmitoleico negli olii. Questo grafico mostra uan peretta separazione degli acidi negli olii provenienti da SA , una leggere separazione degli olii provenienti da NA, mentre per le altre tre regioni di provenienza si nota una sovrapposizione di valori.

3.5 Istogrammi di frequenza

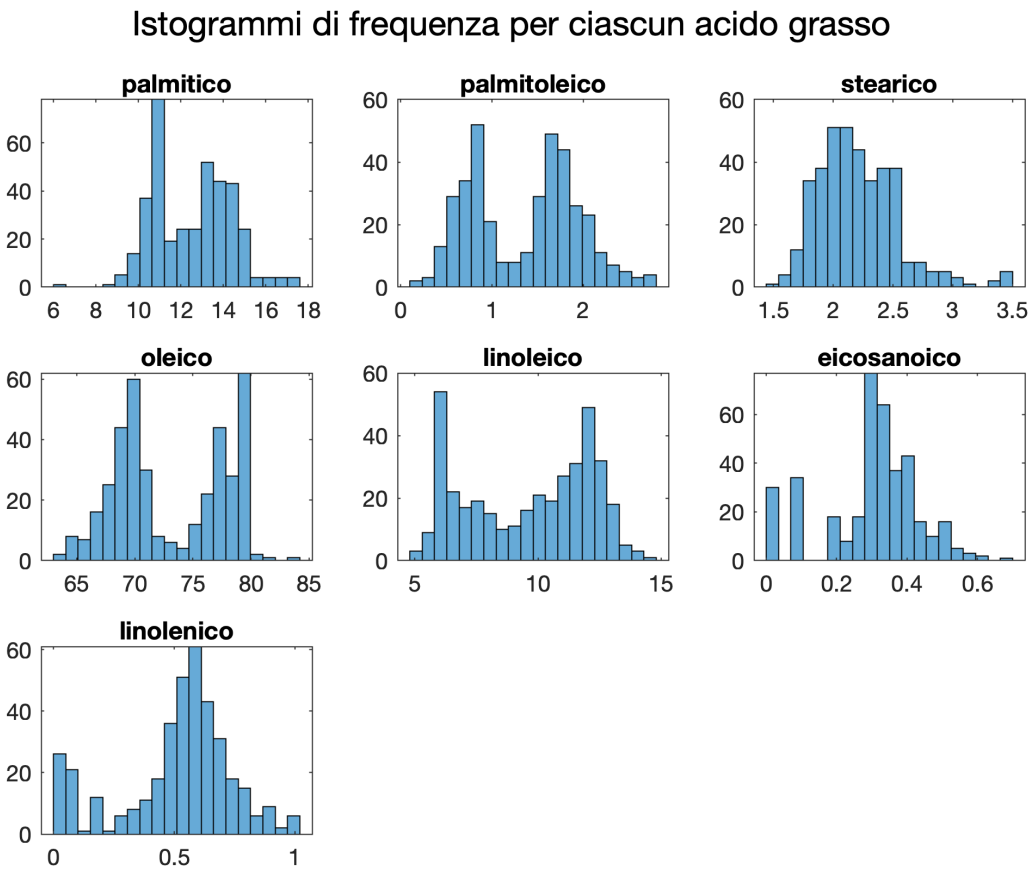


Figure 3.5: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.

Gli istogrammi di frequenza mostrano che la maggior parte degli acidi grassi non segue una distribuzione normale. In particolare, si osservano forme bimodali (quindi con due picchi) (soprattutto per l'acido oleico e palmitico), indicanti la presenza di sottogruppi nei dati riconducibili alle differenti regioni di provenienza. Queste strutture a gruppi suggeriscono differenze nelle regioni di provenienza dell'olio, potremmo quindi usare i due acidi grassi oleico e palmitico per discriminare la regione di appartenenza dell'olio.

3.6 Boxplot per variabile e categoria

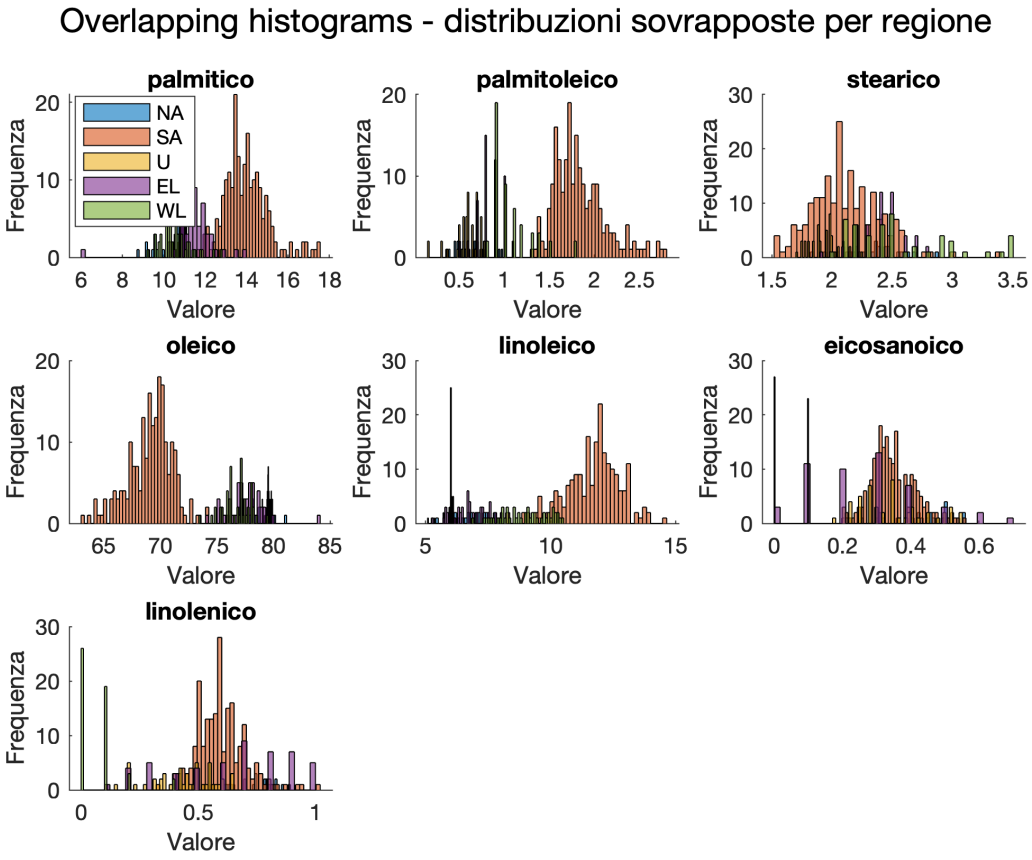
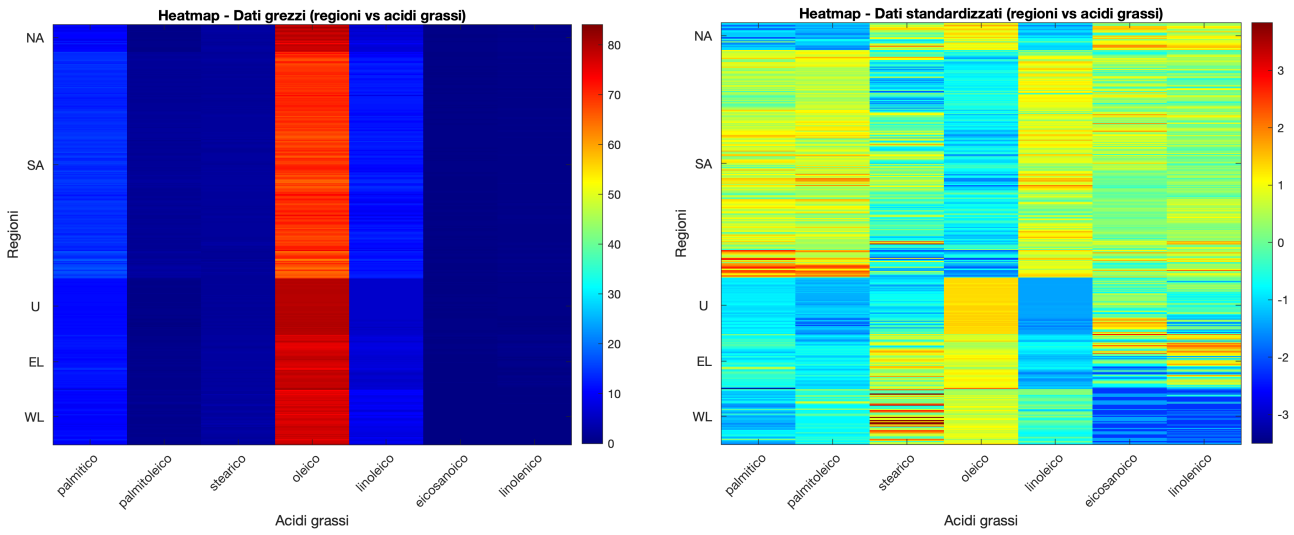


Figure 3.6: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.

Da questo grafico siamo in grado di notare come la distribuzione di ogni acido sia discostata dalle altre distribuzioni per la regione SA. Le variabili, nel nostro caso acidi grssi, che permettono di discriminare meglio in generale le regioni di provenienza sono: Linoleico, Palmitoleico e Eicosanoico. Per gli altri notiamo che vi è una forte sovrapposizione.

3.7 Imagesc (dati grezzi e standardizzati)



(a) Dati grezzi

(b) Dati standardizzati

Figure 3.7: Confronto tra heatmap dei dati grezzi e standardizzati.
Dalla heatmap a dati grezzi possiamo dedurre che l'acido grasso oleico ha un valore molto più alto rispetto agli altri indipendentemente dalla regione di produzione. Mentre per quanto riguarda gli altri acidi, specialmente palmitoleico, stearico, eicosanoico e linoleinico sono, indipendentemenre dalla regione a valori bassissini prossimi allo zero.
Dalla heatmap che riporta invece i dati standardizzati è possibile confutare le assunzioni fatte poco prima infatti vediamo che rimane vera l'assunzione sull'acido olieco che ha valori tendenzialmente opposti rispetto agli altri acidi grassi nelle stesse regioni, ma non mantiene più lo stesso valore da regione a regione. Anche gli altri acidi citati sopra ora sono molto più vari come distribuzione di valori e possiamo dire però che palmitico, palmitoleico e linoleico hanno quasi gli stessi valori nelle varie regioni, mentre eicosanoico ha i medesimi valori dell'acido linolenico.

3.8 Matrice di correlazione

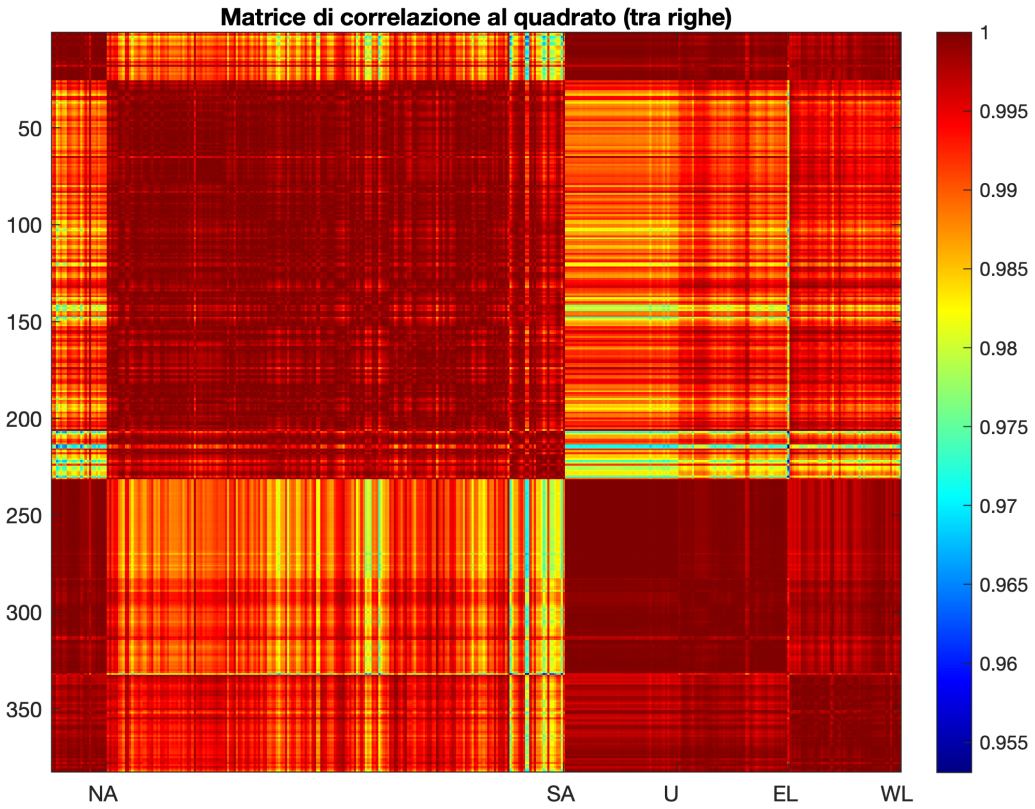


Figure 3.8: Grafico delle coordinate parallele per i campioni di olio extra vergine di oliva.
Da questo plot possiamo notare come la diagonale di correlazione sia, coerentemente con quello che ci aspettiamo, mantenuta. Vi è inoltre una sorta di diagonale secondaria, ovvero gli olii prodoti in NA, in U e in EL risultano abbastanza correlati agli olii prodotti in SA.

4

Conclusioni

Tramite l'opportuna analisi dei grafici è stato possibile notare da ognuno risultati e caratteristiche del dataset diverse. Sono stati infatti ritrovati molteplici trend lineari e inversi, distribuzioni spesso a gruppi e non normali, variabili correlate tra loro e gruppi di olii, che indipendentemente da quale acido grasso stessimo considerando, si distaccava e separava dagli altri gruppi. Si sono analizzate anche differenze nei risultati ottenuti da heatmap con dati grezzi e standardizzati, mostrando come il secondo approccio possa rendere molto più chiari certi pattern che dai dati grezzi non sono osservabili.
Possiamo dunque concludere che per effettuare un analisi esplorativa a 360 gradi, anche per dataset piccoli, è necessario utilizzare multiple tecniche di rappresentazione dei dati e talvolta, anche combinarle tra loro.

Bibliography

[1] The MathWorks, I. (2025). *MATLAB version 9.16.0 (R2025a)*. The MathWorks, Inc., Natick, Massachusetts, United States.